CAHILL GORDON & REINDEL LLP
JOEL KURTZBERG (admitted *pro hac vice*)
FLOYD ABRAMS (admitted *pro hac vice*)
JASON ROZBRUCH (admitted *pro hac vice*)
LISA J. COLE (admitted *pro hac vice*)
32 Old Slip
New York, New York 10005
Telephone: 212-701-3120
Facsimile: 212-269-5420
jkurtzberg@cahill.com

DOWNEY BRAND LLP
WILLIAM R. WARNE (Bar No. 141280)
bwarne@downeybrand.com
MEGHAN M. BAKER (Bar No. 243765)
mbaker@downeybrand.com
621 Capitol Mall, 18th Floor
Sacramento, CA 95814
Telephone: 916-444-1000
Facsimile: 916-520-5910

Attorneys for Plaintiff X Corp.

UNITED STATES DISTRICT COURT

EASTERN DISTRICT OF CALIFORNIA

SACRAMENTO DIVISION

X CORP.,

Plaintiff,

v.

ROBERT A. BONTA, Attorney General of California, in his official capacity,

Defendant.

No. 2:23-cv-01939-WBS-AC

AFFIDAVIT OF [redacted] IN SUPPORT OF X CORP.'S MOTION FOR PRELIMINARY INJUNCTION

[redacted], being duly sworn, deposes and states as follows:

1. I am a [redacted] on the Trust and Safety Team at X Corp, covering product, policy, and operations. In this role, I assist with the creation, implementation, and enforcement of policies — including content moderation policies — on the X platform.

2. I am submitting this affidavit in support of Plaintiff's Motion for Preliminary Injunction. I have personal knowledge of the facts set forth herein, unless otherwise noted. If called upon as a witness, I could and would competently testify to those facts.

I. Applicability of AB 587

3. I have reviewed and am familiar with the law in this case: California's AB 587. AB 587 applies to X Corp. and its X platform.

4. X Corp. is a "social media company," as defined by § 22675(d), as it is a "person or entity that owns or operates a social media platform," as defined by the statute. X Corp. owns and operates X, which is a social media platform, as defined by AB 587.

5. X Corp. generated more than one hundred million dollars in gross revenue during the 2023 calendar year.

6. X is a "social media platform," as defined by § 22675(e), as it is a public internet-based application that has users in California. A substantial function of X is to connect users in order to allow users to interact socially with each other within

the service or application. X allows users to construct public or semipublic profiles for purposes of signing into and using the application. X also allows users to create or post content viewable to others and populate a list of other users with whom an individual shares a social connection within the system.

II. Burdens of AB 587

7. AB 587 purports to require large social media companies, such as X Corp., to (1) post terms of service dictated by the government and include terms about how content is moderated on their platforms (the "Terms of Service Requirement") and (2) submit, on a semi-annual basis, to the California Attorney General a "terms of service report" that includes, among other things, (a) "a detailed description of content moderation practices used by the social media company for that platform"; (b) information about whether, and if so how, the social media company defines and moderates (i) hate speech or racism, (ii) extremism or radicalization, (iii) disinformation or misinformation, (iv) harassment, and (v) foreign political interference; as well as (c) information and statistics about actions taken by the social media company to moderate these categories of content (the "Terms of Service Report"). §§ 22676, 22677.

8. The Terms of Service Requirement in AB 587 burdens X Corp. by letting the State mandate the format and contents of X Corp.'s Terms of Service. Specifically, it requires X Corp. to,

among other things:

a. Make public commitments to respond to and resolve reports of flagged content or violations of the terms of service within a specified time period. § 22676(b)(2).

b. Create and publish Terms of Service in twelve different languages – specifically, the Medi-Cal threshold languages pursuant to the Health and Safety Code § 128552, if the platform offers product features in those languages. § 22676(c). Upon information and belief this includes: (1) Arabic, (2) Armenian, (3) Cambodian, (4) Cantonese, (5) Farsi, (6) Hmong, (7) Korean, (8) Mandarin, (9) Russian, (10) Spanish, (11) Tagalog, and (12) Vietnamese. *See* **Exhibit 1**, which is a true and correct copy of the *Primary Language of Newly Medi-Cal Eligible Individuals*, California Department of Health Care Services, available at https://data.chhs.ca.gov/dataset/primary-language-of-newly-medi-cal-eligible-individuals/resource/706bf0a7-9bb4-4674-9b58-917daac10d25 (last visited Oct. 6, 2023). X Corp. currently offers product features in eight of twelve applicable Medi-Cal threshold languages.

9. Compliance with the Terms of Service Report of AB 587 would be even more burdensome – particularly in light of the quantity of content posted on X.

10. Our current best estimate is that, over the past 30 days, there have been on average, approximately 7,000 posts (formerly called "tweets") made on X every second. If we extrapolate from that number, that would be approximately 420,000 posts per minute, 604.8 million posts per day, and almost 221 billion posts per year.

11. To complete the Terms of Service Report required by AB 587, X Corp. would need to keep track of and categorize any and all of the content moderation decisions made as to the almost 221 billion posts each year, including the millions of moderation actions taken each year by automated enforcement tools.

12. The Terms of Service Report required by AB 587 would force X Corp. to disclose "any existing policies intended to address" the categories of content identified by the statute (i.e., "hate speech or racism," "extremism or radicalization," "disinformation or misinformation," "harassment," and "foreign political interference." § 22677(a)(4)(A) and (a)(3).

13. This would be extremely difficult for X Corp. to do. X Corp. already has detailed descriptions of its content moderation policies available to the public online. (This is addressed in more detail below.) But X Corp.'s categories of content moderation, while comprehensive, do not align with the categories

identified in the statute. And the listed statutory categories are controversial and difficult to define.

14. For example, X Corp. does not currently regulate "hate speech," "racism," or "extremism" *per se*, yet these are three categories of conduct that AB 587 forces social media companies to publicly address. § 22677(a)(3). But X Corp. regulates "violent speech," "hateful conduct," and "violent and hateful entities," categories of content that may include content that some people might argue constitutes "hate speech," "racism," and/or "extremism." *See* **Exhibit 2** (Violent Speech Policy), **Exhibit 3** (Hateful Conduct Policy) and **Exhibit 5** (Violent and Hateful Entities Policy). Thus, X Corp.'s current policies do not fit within the allotted statutory categories. And it is not at all clear whether X Corp.'s "violent speech," "hateful conduct," and "violent and hateful entities" policies are "intended to address" the statutory categories, which tend to mean different things to different people.

15. Because the statute provides the Attorney General with discretion to impose significant civil penalties for noncompliance – of up to $15,000 per violation per day, § 22678 – which includes a violation for material omissions or misrepresentations in the Terms of Service Report, § 22678 at (a)(2)(C), it gives the Attorney General complete discretion to determine whether he believes that it would be a violation to submit a report that did not include

information about a policy like the "hateful conduct" policy, which does not fit neatly into the statutory categories.

16. The Terms of Service Report required by AB 587 would also force X Corp. to disclose "[h]ow automated content moderation systems enforce terms of service and when these systems involve human review." § 22677(a)(4)(B).

17. This would harm X Corp. because it would force X Corp. to disclose publicly information about its automated content moderation systems that is highly confidential and that could put X Corp. at a competitive disadvantage if made public. Moreover, disclosure of this information to the public would undermine X Corp.'s policy enforcement by providing malicious actors with information that they would likely leverage in attempts to circumvent and manipulate our policy enforcement mechanisms. Disclosure of this information is therefore likely to compromise the safety and integrity of the X platform.

18. Perhaps most significantly, the Terms of Service Report required by AB 587 would also force X Corp. to:

a. Collect, analyze, and generate reports detailing (i) the total number of flagged items of content; (ii) the total number of actioned items of content; (iii) the total number of actioned items of content that resulted in action taken by the social media company against the user or group of users responsible for the

content; (iv) the total number of actioned items of content that were removed, demonetized, or deprioritized by the social media company; (v) the number of times actioned items of content were viewed by users; and (vi) the number of times actioned items of content were shared, and the number of users that viewed the content before it was actioned. § 22677(a)(5)(A).

b. All of the information in ¶ 14(i) above must then be disaggregated into (i) the category of content; (ii) the type of content, including, but not limited to, posts, comments, messages, profiles of users, or groups of users; (iii) the type of media of the content, including, but not limited to, text, images, and videos; (iv) how the content was flagged, including, but not limited to, flagged by company employees or contractors, flagged by artificial intelligence software, flagged by community moderators, flagged by civil society partners, and flagged by users; (v) how the content was actioned, including, but not limited to, actioned by company employees or contractors, actioned by artificial intelligence software, actioned by community moderators, actioned by civil society partners, and

actioned by users; and (vii) the number of times users appealed social media company actions taken on that platform and the number of reversals of social media company actions on appeal disaggregated by each type of action. § 22677(a)(5)(B).

19. The reporting requirement under AB 587 requires three reports be submitted to the Attorney General in 2024. The first report is due on January 1, 2024 and covers activity within the third quarter of 2023. § 22677(b)(2).

20. It would be enormously burdensome to create and categorize those records for the almost 221 billion posts made on X each year. X Corp. does not currently have the tools, infrastructure, or staff levels necessary to meet the onerous reporting requirements of AB 587. Indeed, X Corp. would need to design, build, and implement entirely new tools and workflows, including a new categorization system for moderation actions, in order to comply in good faith with the requirements of AB 587.

21. In the third quarter of 2023, there were approximately 55 billion posts on X. X Corp. has not undertaken the burden of identifying how many of these 55 billion posts may constitute hate speech, racism, extremism, radicalization, disinformation, misinformation, harassment, or foreign political interference, and doing so would require making numerous highly controversial decisions regarding what posts fit in each of those categories.

Nor has X Corp. aggregated statistics on actions and appeals taken with respect to this content, as required in § 22677(a)(5).

22. My team's initial estimate is that implementing the infrastructure and processes necessary to comply with the requirements of AB 587 would require at least six months and involve at least thirty X Corp. employees — diverting engineering, business, and legal resources away from existing, mission-critical projects. To comply with AB 587, X Corp. would need to indefinitely commit resources to the maintenance and operation of this new compliance infrastructure.

23. X Corp. would need to hire new employees and/or onboard contractors in order to allocate resources to achieve good faith compliance with AB 587. Doing so would cost X Corp. hundreds of thousands or millions of dollars per year.

24. This does not even factor in burdens imposed by follow-up questions that will almost surely be asked by the Attorney General about compliance. Given the broad enforcement powers granted to the Attorney General under California law, AB 587 would authorize the Attorney General to issue document demands and follow-up requests to social media companies about their content moderation policies and practices to determine if they have complied with the statute in "reasonable, good faith." § 22678(a)(3).. Responding to such requests would impose significant additional burdens on X Corp. and other social media

companies.

III. The Controversial Nature of Content Moderation

25. Social media content moderation is an inherently controversial undertaking. How social media companies define categories of speech that will or will not be permitted on their platforms and how they apply their rules on content moderation to particular posts are fraught with sensitive and controversial questions and are subject to rigorous debate and controversy. Put another way, deciding what content should appear on a social media platform is a question that engenders considerable debate among reasonable people about where to draw the correct proverbial line.

26. At X Corp., we take very seriously our responsibility to make well-thought out content moderation decisions and to make sure those decisions are accessible and understandable to our users. But in making these difficult calls, we have learned that, no matter what we do, some portion of the public will likely take issue with the way we have made these difficult judgment calls.

27. As the California Assembly Committee on Privacy and Consumer Protection correctly explained in analyzing AB 587, content moderation by social media companies presents those companies with a "complex dilemma" because, whatever those companies do, their decisions are subject to controversy and "both action and inaction by these companies seem[] to be equally maligned[.]" Affidavit of Joel Kurtzberg in Support of Motion for

Preliminary Injunction, Exs. 5, 8, and 9.

28. This is consistent with our experience at X Corp. There is intense public debate and controversy about how to define the categories of content that should be limited on the social media platform and how to apply those categories to content on the social media platform. No matter what decisions are made, there are almost always large groups of people who disagree with them.

29. And the categories identified by the statute are among the most difficult to define and the most controversial to apply. While there is a general consensus that clear and extreme illegal content, such as child pornography, should not be permitted on a social media platform, there is no such consensus as to the categories identified by the statute – e.g., "hate speech, racism, extremism, misinformation, political interference, and harassment." Those categories of content are defined differently by different people and are therefore highly controversial to define and apply in practice.

IV. Content Moderation at X Corp.

30. X Corp. goes to great lengths to make its content moderation policies relevant and transparent. X Corp.'s rules, policies, and procedures about content moderation are all publicly available and are regularly revisited and updated based on X Corp.'s editorial judgments about what should and should not be permitted on the X platform.

31. Annexed hereto are true and correct copies of current X Corp. policies that concern content moderation and transparency efforts:

a. **Exhibit 2:** *Violent Speech Policy*, X Corp., June 2023, available at https://help.twitter.com/en/rules-and-policies/violent-speech (last visited Oct. 6, 2023);

b. **Exhibit 3:** *Abuse and Harassment*, X Corp., June 2023, available at https://help.twitter.com/en/rules-and-policies/abusive-behavior (last visited Oct. 6, 2023);

c. **Exhibit 4:** *Hateful Conduct Policy*, X Corp., Apr. 2023, available at https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (last visited Oct. 6, 2023);

d. **Exhibit 5:** *Violent and Hateful Entities Policy*, X Corp., Apr. 2023, available at https://help.twitter.com/en/rules-and-policies/violent-entities (last visited Oct. 6, 2023);

e. **Exhibit 6:** *Abusive Profile Information*, X Corp., available at https://help.twitter.com/en/rules-and-policies/abusive-profile (last visited Oct. 6, 2023);

f. **Exhibit 7:** *Crisis Misinformation Policy*, X Corp., Aug. 2022, available at https://help.twitter.com/en/rules-and-policies/crisis-misinformation (last visited Oct. 6,

2023);

g. **Exhibit 8:** *Synthetic and Manipulated Media Policy*, X Corp., Apr. 2023, available at https://help.twitter.com/en/rules-and-policies/manipulated-media (last visited Oct. 6, 2023);

h. **Exhibit 9:** *Civic Integrity Policy*, X Corp., Aug. 2023, available at https://help.twitter.com/en/rules-and-policies/election-integrity-policy (last visited Oct. 6, 2023).

32. Additional X Corp. rules, policies, and procedures that may address content moderation and transparency efforts are available to the public at http://help.twitter.com/en/rules-and-policies (last visited Oct. 6, 2023).

33. As detailed in these policies, X Corp. does moderate content that may arguably be covered by some of the controversial categories set forth in AB 587. For example, X Corp. moderates hateful conduct, crisis misinformation, violent speech, abuse and harassment, child sexual exploitation, abusive profiles, violent and hateful entities, and glorification of violence.

34. X Corp. crafted its rules, policies, and procedures with the objective of ensuring all can engage in the critical public debates surrounding these topics freely and safely.

35. In creating its rules, policies, and procedures around

content moderation, X Corp. takes into consideration insights gained over the past seventeen years that the platform has been operational, including feedback from trusted experts at X Corp., our users, and the public at large.

36. X Corp. dedicates immense time, energy, and resources into crafting these rules, policies, and procedures and ensuring they are accessible and understandable to its users.

37. Through AB 587, the government is impermissibly trying to frame the content moderation debate and force social media companies like X Corp. to conform moderation practices to the government's content moderation priorities and categorizations.

38. By compelling speech about these controversial topics, the State is, in our view, trying to pressure X Corp. to regulate content in the way it wants by framing the public debate about content moderation. If X Corp. does not regulate the categories of content identified by the statute, then the State and/or the public may easily try to pressure X Corp. to do so by drawing attention to that fact or threatening enforcement actions unless X Corp. agrees to do so.

39. X Corp. would be harmed by AB 587's requirements that it make certain statements about a controversial topic against its will.

Dated: [redacted]
October 6, 2023

[redacted]

Sworn to before me this
6 day of October, 2023

See attached Certificate
Notary Public

A notary public or other officer completing this certificate verifies only the identity of the individual who signed the document to which this certificate is attached, and not the truthfulness, accuracy, or validity of that document.

[redacted] } SS.

Subscribed and sworn to (or affirmed) before me on this 6 day of October, 2023, by

[redacted], proved to me on the basis of satisfactory evidence

to be the person(s) who appeared before me.

[redacted] CLAIRE GILBERT
COMM.# 2432385
NOTARY PUBLIC - [redacted] COUNTY
MY COMM. EXP. DEC. 25, 2026

NOTARY'S SIGNATURE

PLACE NOTARY SEAL IN ABOVE SPACE

OPTIONAL INFORMATION

The information below is optional. However, it may prove valuable and could prevent fraudulent attachment of this form to an unauthorized document.

CAPACITY CLAIMED BY SIGNER (PRINCIPAL)

- [x] INDIVIDUAL
- [ ] CORPORATE OFFICER _____ TITLE(S)
- [ ] PARTNER(S)
- [ ] ATTORNEY-IN-FACT
- [ ] TRUSTEE(S)
- [ ] GUARDIAN/CONSERVATOR
- [ ] OTHER: _____

DESCRIPTION OF ATTACHED DOCUMENT

[redacted]

16
NUMBER OF PAGES

10/6/23
DATE OF DOCUMENT

OTHER

ABSENT SIGNER (PRINCIPAL) IS REPRESENTING:
NAME OF PERSON(S) OR ENTITY(IES)

RIGHT THUMBPRINT OF SIGNER

Top of thumbprint here